Description

In the alien language classification task, participants have to train a model on text samples collected by our deep space satellite from three alien races living on different planets. Our satellite was able to record and automatically transcribe several text samples, but the recordings are mixed together. After working continuously for 5 years, a team of British scientists was able to determine the source planet for a bunch of training samples, even if the text appears to look like gibberish. Therefore, participants are asked to help our team of British scientists by building a machine learning model to assign the correct labels for the remaining test samples.

Evaluation

The evaluation measure is the macro F1 score computed on the test set. The macro F1 score is given by the mean of the F1 scores computed for each class. The F1 score is given by:

F1=2⋅P⋅RP+R,

where P is the precision and R is the recall.

Public versus Private Test Evaluation

The public leaderboard is calculated on approximately 20% of the test data. These examples are randomly chosen. The final results will be based on the other 80%, so the final standings may be different. In this context, you will have to choose 2 submissions that you think will attain the best performance on the 80% of the test data that is not used for the public leaderboard.

Submission Format

For every sample in the dataset, submission files should contain two columns: id and label. The id is the identifier associated to a data sample. The label should be the class label (1 2, or 3) predicted for the corresponding data sample.

The file should contain a header and have the following format:

id,label  
1,1  
2,1  
...

## Task

Participants have to train a model on alien text samples. Therefore, participants have to build a model for an alien language classification task, in which a classification model is required to discriminate between languages spoken by aliens on three different planets: Bellerophon (label 1), Ymir (label 2), Nanook (label 3).

The training data is composed of 15,000 samples. The validation set is composed of 5,000 samples.

## File Descriptions

* **train\_samples.txt** - the training data samples (one sample per row)
* **train\_labels.txt** - the training labels (one label per row)
* **validation\_samples.txt** - the validation data samples (one sample per row)
* **validation\_labels.txt** - the validation labels (one label per row)
* **test\_samples.txt** - the test data samples (one sample per row)
* **sample\_submission.txt** - a sample submission file in the correct format

## Data Format

### **Samples File**

The data samples are provided in the following format based on TAB separated values:

108345 Uq$%y gkuKDuZ\*KmH quf& qKf Du\*& Du\*ZX; Kf DY\*fHm

101973 YcîT » ? \*Y\*qK# Yuț’;Du\*ăuH H ènSè ènSè KièâT Z\*\*;

Each line represents a data sample where:

* The first column shows the ID of the data sample.
* The second column is the actual data sample.

### **Labels File**

The labels are provided in the following format based on TAB separated values:

108345 1

101973 1

Each line represents a label associated to a data sample where:

* The first column shows the ID of the data sample.
* The second column is the actual label.